

Extracting Events from Clinical Text Using Natural Language Processing

Shan Rajapakshe

Supervisor: Dr. Ramon Lawrence

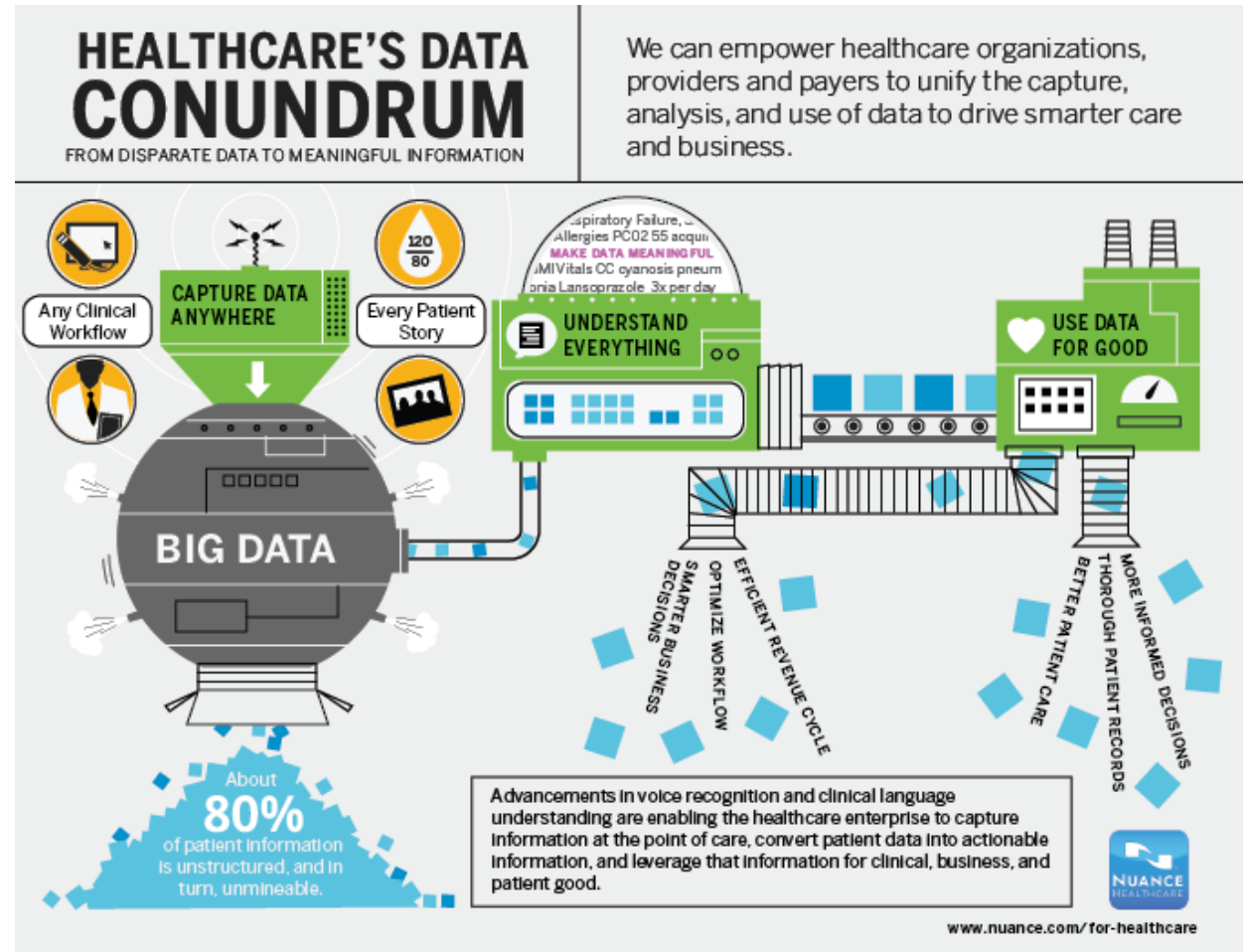
Computer Science

Overview

- Background
 - Healthcare data
 - Information Extraction
 - Dataset
- My Work
 - Clinical Events
 - Features
 - Algorithms
- Results
 - Evaluation
- Moving Forward

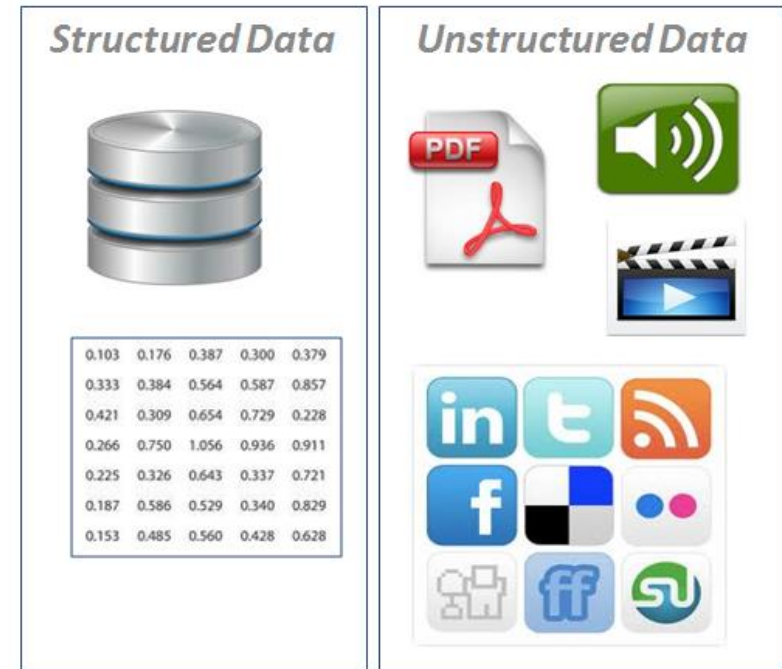
Healthcare Has a Lot of Data

- Pharmacy
- GP visits
- Lab Reports
- Pathology
- Imaging
- **Progress Notes**



Nature of Healthcare Data

- Mix of both unstructured and structured data
- Structured data:
 - Easy to feed into a computer
 - I.e. Data in a spreadsheet
- Unstructured data:
 - Much messier
 - Harder to represent and 'understand'
 - Images, sounds, video, **natural language text**



My Goal – Solve a Subproblem

- Extract relevant **events** from unstructured **clinical text**

Why?

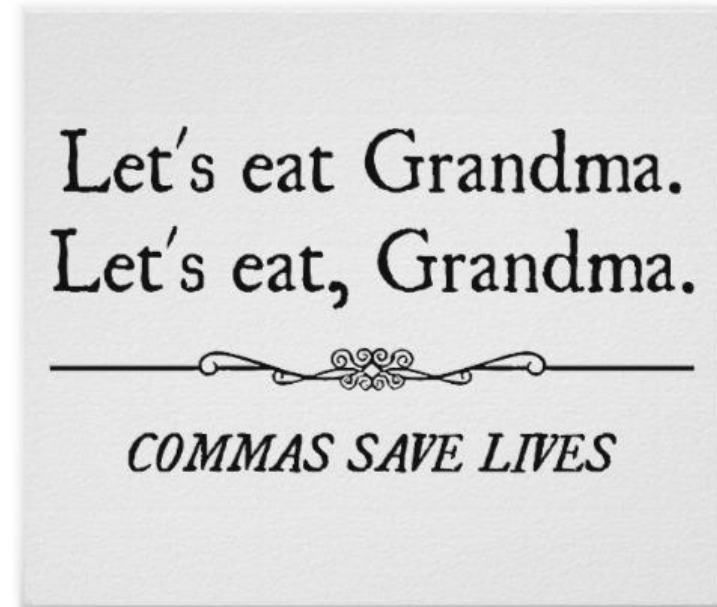
- Currently Physicians need to look back on data
- Less time reading charts, **more time with patients!**



What is Event Extraction?

- "...Extraction of complex combinations of relations between actors (entities), performed after executing a series of Natural Language Processing steps..."
- I.e. Finding a relevant event within text data

Difficult because of **ambiguity**



Methods of Information Extraction

- Expert Systems
 - Leverage pre-existing knowledge
 - Often use patterns or rules
 - Limited by scope of knowledge
- Data Driven
 - Use features from text
 - Apply statistical methods and Machine learning
 - Limited by the data

Dataset

- Clinical Progress notes from a set of 262 Lung Cancer Patients from the BC Cancer Agency
- Had to be anonymized (no patient identifiable information left, but able to access the specific patient if necessary)
 - ~ 10 Charts per patient
 - ~ 2700 Charts total
 - ~ 56000 sentences

My Goal – Revisited

- Extract relevant **events** from unstructured **clinical text**

My Goal – Revisited

- Extract relevant **events** from unstructured **clinical text** using a **data driven approach**

My Goal – Revisited

- Extract relevant **events** from unstructured **clinical text** using a **data driven approach**
 - Determine relevant events
 - Prepare data
 - Find appropriate features
 - Choose suitable machine learning methods
 - Apply features to methods
 - Evaluate results

Tools

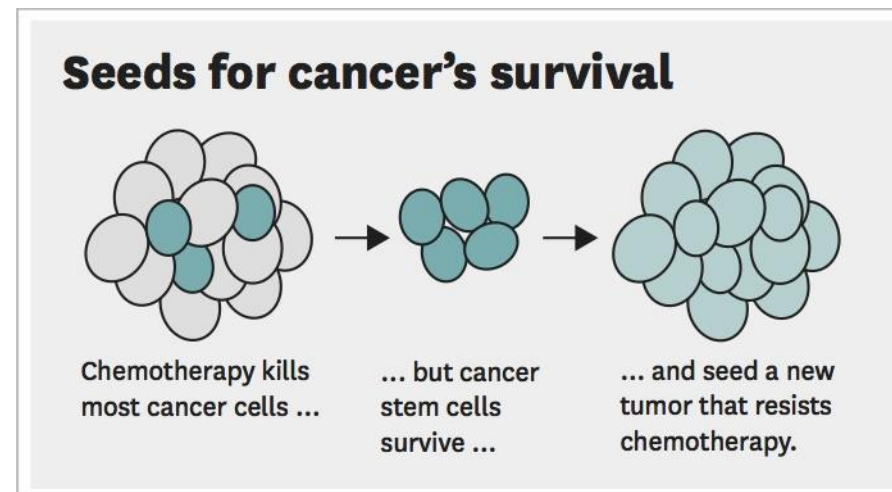


As well as NLTK, Numpy, Faker, and others!

Clinically Relevant Events

Had to define what were important events

- Treatments
 - Chemotherapy
 - Radiation
 - Surgery
 - Palliative
- Recurrence



Data Preparation

- Pull data from CAIS (at BCCA)
- Convert PDFs to text
- Anonymize and obfuscate patient information

Recurrence	Chemotherapy	Radiation	Surgery	Palliative
0	1	0	0	0
0	0	0	0	0
0	0	0	0	1
0	1	0	0	0
-1	0	0	0	0
0	0	0	0	0
0	0	0	1	0
1	0	0	0	0

Features

- How to let a computer represent textual data?
- Balance between understandability for humans vs ease of use for machine

Lorem ipsum dolor sit amet, consectetur
adipiscing elit, sed do eiusmod tempor
incididunt ut labore et dolore magna
aliqua. Ut enim ad minim veniam, quis
nostrud exercitation ullamco laboris nisi
...



```
0010100101 01001 01001 101 1010  
11001 10101010 1 1001010101011  
0101 01 10 101 011 0101 10 101  
10101 10 1010 101010 101 101010101  
1010101 101101 100001 01010 0100101  
...
```

Features Used

Bag of Words representations

- 1-grams
- 2-grams
- TFDIF

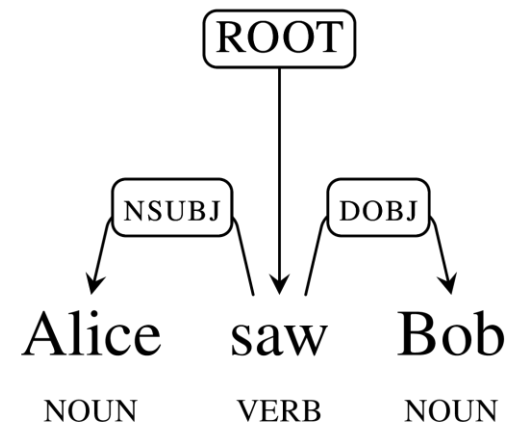
Alice saw Bob
Susan called Bob



Alice
Bob
Saw
Susan
Called

Parts of Speech Tags

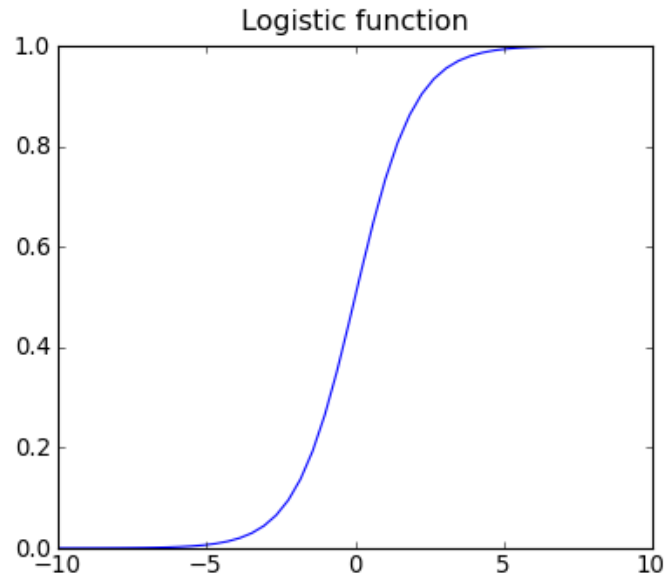
- Dependency Parse Trees
- Named Entities



Algorithms

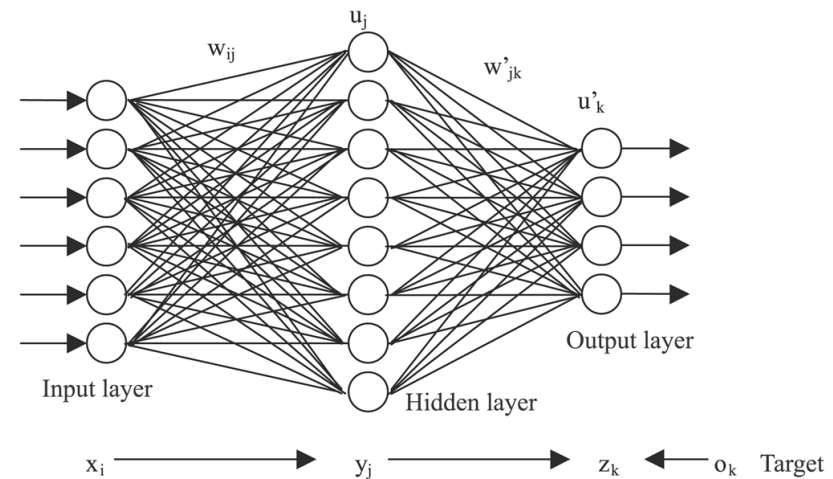
Logistic Regression

- Robust
- Quick
- powerful

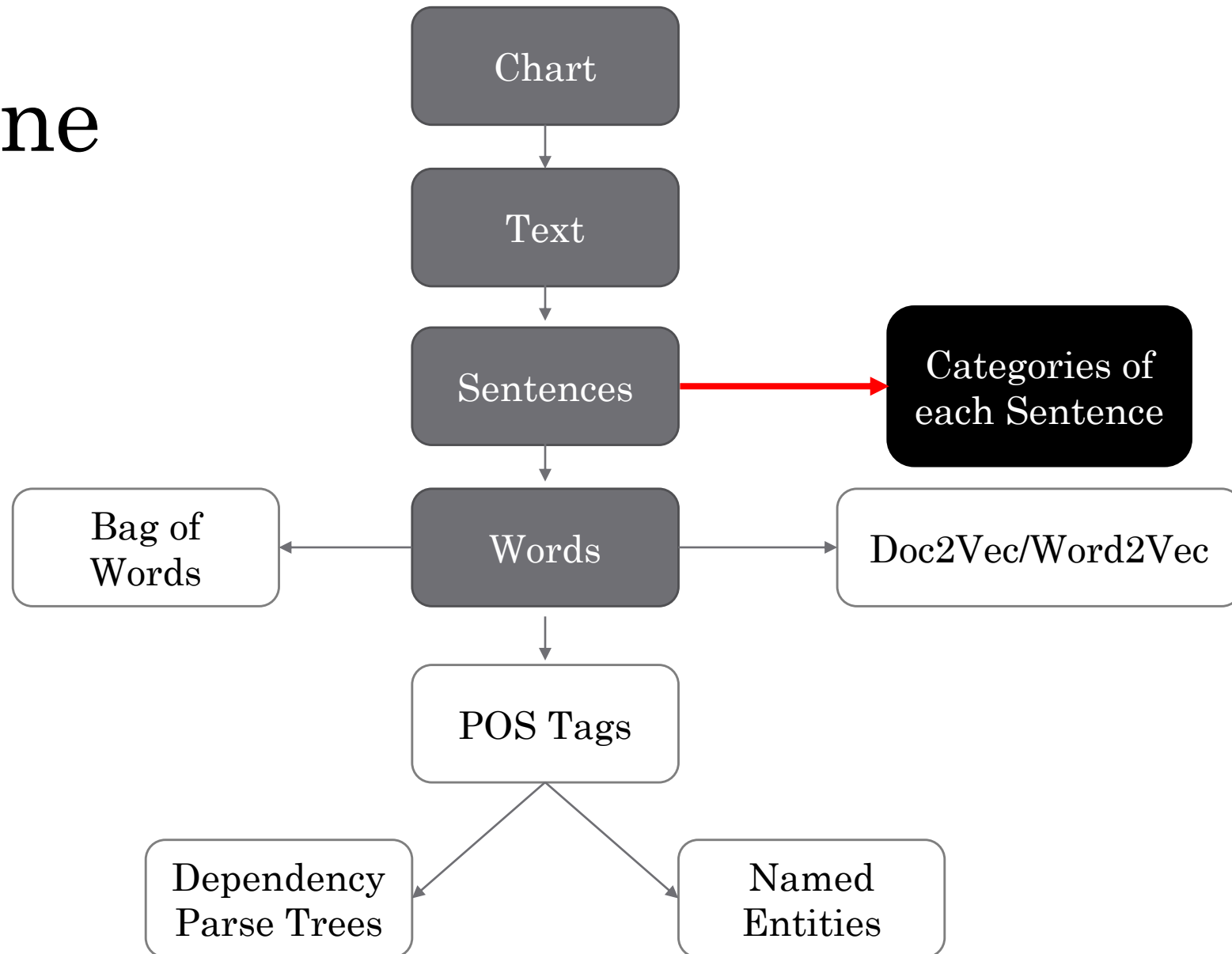


Multilayer Perceptron

- Slow
- Requires lots of data
- Lots of potential



Pipeline

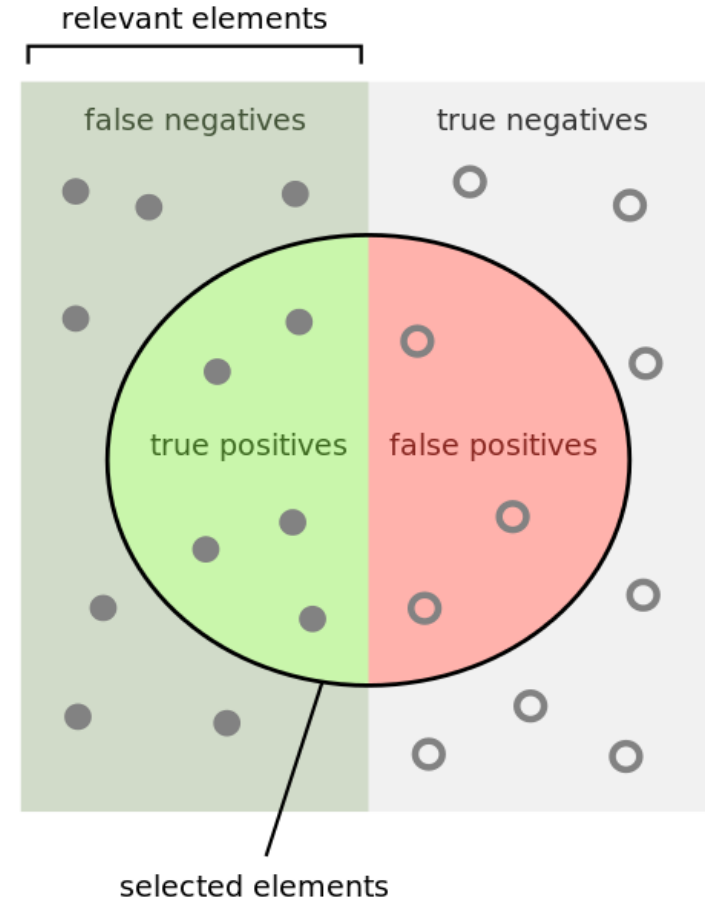


Evaluation

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



Preliminary Results

- Using 3-Gram BOWs for Recurrence Events

---- LOGISTIC REGRESSION ----

Normalized Score: 0.9860

Raw Score: 2329

---- PERCEPTRON ----

Normalized Score: 0.9814

Raw Score: 2318

	Precision	Recall	F1-Score	Support
-1	0.68	0.73	0.70	26
0	0.99	0.99	0.99	2308
1	0.6	0.54	0.57	28
Avg / Total	0.99	0.99	0.99	2362

	Precision	Recall	F1-Score	Support
-1	0.46	0.73	0.57	26
0	0.99	0.99	0.99	2308
1	0.43	0.11	0.17	28
Avg / Total	0.99	0.99	0.99	2362

Preliminary Results

- Using 3-Gram BOWs for Recurrence Events

---- LOGISTIC REGRESSION ----

Normalized Score: 0.9860

Raw Score: 2329

---- PERCEPTRON ----

Normalized Score: 0.9814

Raw Score: 2318

	Precision	Recall	F1-Score	Support
-1	0.68	0.73	0.70	26
0	0.99	0.99	0.99	2308
1	0.6	0.54	0.57	28
Avg / Total	0.99	0.99	0.99	2362

	Precision	Recall	F1-Score	Support
-1	0.46	0.73	0.57	26
0	0.99	0.99	0.99	2308
1	0.43	0.11	0.17	28
Avg / Total	0.99	0.99	0.99	2362

What does this mean?

- Majority of sentences **not relevant to these events**
- Potential to **streamline physician workflow** by only showing relevant information

---- LOGISTIC REGRESSION ----

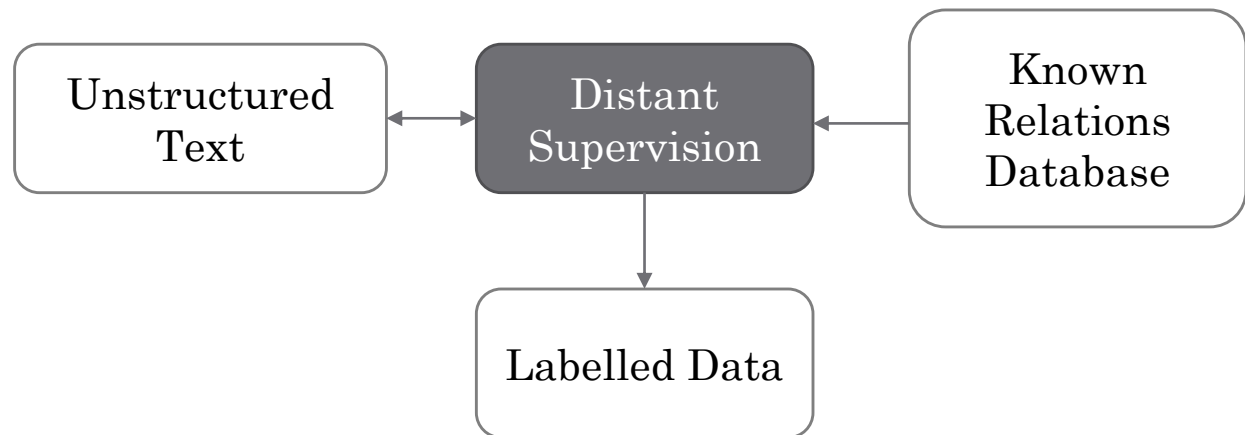
	Precision	Recall	F1-Score	Support
-1	0.68	0.73	0.70	26
0	0.99	0.99	0.99	2308
1	0.6	0.54	0.57	28
Avg / Total	0.99	0.99	0.99	2362

Conclusions

- Open Source NLP tools **can be used to extract clinical events** from healthcare data
- **Simple algorithms** work well for small datasets
- Methods work better than a naïve classifier
- Data cleaning is **hard**

Going Forward

- Already sharing data with group in Victoria
- Improve labelling/data generation process
- Find a meaningful way to represent information to physicians



Thanks!

- To Dr. Ramon Lawrence for providing guidance and asking me tough questions
- To Dr. Jonn Wu for the space to work at the BCCA and the idea
- To Dr. Cheryl Ho for the patient cohort

And thank you!

Questions?